LAMP-TR-104 UMIACS-TR-2003-78
CAR-TR-986
CS-TR-4510

USE OF OCR FOR RAPID CONSTRUCTION OF BILINGUAL LEXICONS

Burcu Karagol-Ayan, David Doermann, and Bonnie J. Dorr

Language and Media Processing Laboratory
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
{*burcu,doermann,bonnie*}*@umiacs.umd.edu*

Abstract

This paper describes an approach to analyzing the lexical structure of OCRed bilingual dictionaries to construct resources suited for machine translation of low-density languages, where online resources are limited. A rule-based and an HMM-based method are used for rapid construction of MT lexicons based on systematic structural clues provided in the original dictionary. We evaluate the effectiveness of our techniques, concluding that: (1) the rule-based method performs better on dictionaries with a simple structure; (2) the stochastic method performs better on dictionaries with an enriched structure; (3) regardless of the degree of dictionary richness, the rule-based method gives better results for *phrasal* entries than for *single-word* entries; and (4) Our resulting bilingual lexicons are comprehensive enough to provide reasonable MT results when compared to human-constructed lexicons.

Keywords: Bilingual dictinary, OCR, Lexicon

The support of this research by DARPA under contract N66001002891002 and by NSF CISE Research Infrastructure Award EIA0130422 are gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Use of OCR for Rapid Construction of Bilingual Lexicons	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**17**	

1 Introduction

An important requirement for machine translation (MT) is the existence of a bilingual lexicons containing large sets source-language/target-language correspondences. Several researchers have noted that, even for *monolingual* entries, the average time needed to construct a single entry can be as much as 30 minutes (see, e.g., [6, 17, 26]). The construction of *bilingual* entries is even more complicated in that it requires native-speaker knowledge in both languages [3, 4, 18]. Thus, automation of the bilingual lexical acquisition process is a necessity for multilingual processing of any kind.

The wide availability of new electronic resources to NLP researchers has facilitated automated acquisition of bilingual lexicons. Previous approaches to bilingual-lexicon acquisition have involved (1) parallel corpora [9, 15, 21, 23]; (2) comparable corpora [8]; and (3) multilingual thesauri [25]. The reliance on such resources has constrained the application of these approaches to languages that are most frequently used in MT and cross-language information retrieval (CLIR) tasks, e.g., English, French, Spanish, and Chinese. The same approaches are difficult to apply to language pairs involving low-density languages (e.g., Arabic, Cebuano, Turkish) where there are not enough parallel or comparable resources to produce full bilingual lexicons.

This paper describes implemented methods for resource acquisition from *printed* bilingual dictionaries, especially for low-density languages. The basic motivation behind this work is that many languages have printed bilingual dictionaries mapping a low-density language to a high-density language such as English. Ultimately the objective is to discover all supplemental entry-level components of information provided in bilingual dictionaries, e.g., parts of speech, pronunciation, and usage examples. The speed of our lexical-acquisition approach is a unique feature of our work: we aim to generate an online bilingual lexicon very quickly (at most, in a few days).

Our focus is on an implemented *entry-tagging* module for online lexicon construction. We adopt three different methods: rule-based, stochastic, and post-processed stochastic. All utilize the repeating structure of the dictionaries to identify and label the different information types. Human assistance—required for all three techniques—is held to a

minimum. We demonstrate that, whereas the rule-based tagging method performs better on dictionaries in which font is not a distinguishing feature for determining information types, the stochastic tagging method generally performs better on dictionaries in which font is an important feature. We also show that, a post-processing stochastic method improves the results of the stochastic method on *phrasal* entries. Finally, we show that our resulting bilingual lexicons are comprehensive enough to provide the basis for reasonable translation results when compared to human translations.

The next section discusses work related to our approach. In Section 3 we describe our three methods. Section 4 presents our experiments and discusses our results. We conclude with future work.

2 Related Work

In recent years, researchers have become increasingly interested in information extraction from structured printed documents. A key component of their solution is the use of textual features to perform labeling within a block according to some implicit or explicit model. Automatic identification of structural features in OCRed documents has been implemented in approaches where documents are tagged iteratively, using a Standard Generalized Markup Language (SGML) [19]. Such approaches produce a SGML document that can be easily parsed.

In other approaches [13], automatic bilingual-dictionary extraction has relied on stochastic language models based on manually created context-free grammars (CFG) and dictionary-specific stochastic production rules. These approaches are reasonable for dictionaries with a simple structure, e.g., where font is not used to indicate information types. In the general case, however, manual, grammar-based approaches will not be able to handle uncertainty in OCR, and errors in the document analysis.

Our source document is also a bilingual dictionary. However, our approach is designed to tackle some of the issues that hamper approaches based strictly on formal grammars, in particular: (1) complexity and variations within dictionary entries; and (2) noise introduced by OCR and subsequent feature extraction.

3 Approach

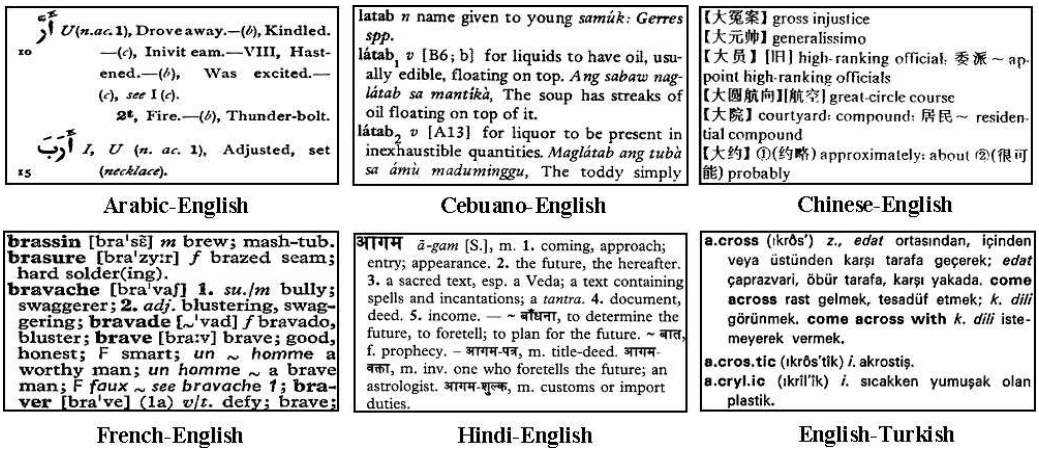

Figure 1: Examples of bilingual dictionaries

We have built an entry-tagging system that can be adapted to different bilingual dictionary formats as well as different languages. Figure 1 illustrates that dictionary formats vary from simple term and phrase translation pairs to full descriptions that contain several different *information types*, i.e., identifiable "chunks" of information associated with bilingual lexical entries.[1]

We borrow a pre-existing text segmenter/analyzer [12],[2] using its output to identify different information types (parts of speech, pronunciation, usage examples) for each bilingual dictionary entry. Table 1 provides a list of potential information types. Two types that we will refer to frequently in this paper are: (1) *Headword*, which refers to the main word that defines the entry; and (2) *Derived Word*, which refers to a word that is lexically related to the headword (e.g., an adjectival form of a verb entry). This list was obtained through manual examination of printed dictionaries.

[1] Although we focus on dictionaries mapping from a low-density language to a high-density language, we have applied our system more broadly, to dictionaries that contain the reverse mapping. This is important for cases where printed resources are limited to the less preferred bilingual direction. We expect the output of such an analysis to be easily inverted using standard dictionary-inversion techniques [16].

[2] The pre-existing text segmenter/analyzer was induced through standard image pre-processing and machine-learning techniques: (1) The printed dictionary pages were scanned and divided into logical entries containing words and their associated layout features (the font or color used, the location of the word on the page, etc); (2) The layout features were then used as input to a machine-learning algorithm to bootstrap a customized segmenter/analyzer.

Headword	Translation
Pronunciation	Tense
Part of speech (POS)	Gender
Plural Form	Number
Domain	Context
Cross reference	Language
Antonym	Derived word
Synonym	Derived word translation
Inflected form	Usage Example
Irregular form	Usage Example translation
Alternative spelling	Idiom
Explanation	Idiom translation

Table 1: Information Types Found in Bilingual Dictionaries

We assume that the OCRed and pre-segmented dictionaries provide the following information as input to our entry-tagging system:

- each page is divided into dictionary entries
- each entry is associated with an entry type
- for each entry, lines and tokens are identified
- for each token, font style is provided

where a *token* is a set of glyphs (i.e., a visual representation of a set of characters) in the OCRed output, separated by white space. Given an input in this format, our entry-tagging system associates labels with each information type provided by a token or group of tokens in the entry. The system requires input from a human operator who is familiar with, but not necessarily expert in, the language of interest.

Publishers of dictionaries typically use a combination of methods to impose structure on lexical entries. Functional properties (changes in font, font style, font-size, etc.) make the information type implicit, keywords provide an explicit interpretation of the information type, and various separators impose an overall structure on the entry. For instance, a boldface font may indicate headwords, italics may indicate usage examples, keywords may designate the POS, commas may be used to separate different translations, and a numbering system may be used to identify different senses of the word. Our system uses these clues to identify information types associated with a token (or group of tokens)

in a lexical entry.

We have implemented three different methods for entry tagging: a rule-based model, a stochastic Hidden Markov (HMM) model, and a post-processed stochastic HMM model. One of the challenges we faced was the handling of noisy input provided by the pre-existing OCR/segmenter. The rule-based method accommodates noise by allowing for a relaxed matching of OCRed output to information types. The HMM method and post-processed stochastic HMM method are *inherently* noise-tolerant due to the statistical nature of the training procedure underlying the models.

The overall architecture is shown in Figure 2. We now describe each of these three methods in detail.

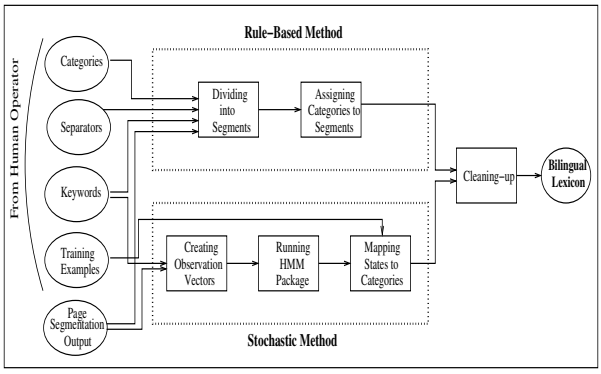

Figure 2: Overall entry tagging design

3.1 Rule-Based Method

Our rule-based tagging approach uses the functional properties of tokens and their relationships to each other in order to assign labels to each information type in a dictionary entry. Rule-based tagging utilizes three different types of clues—font style, keywords and separators—to tag the entries in a systematic way. The key is to discover the regularities in the occurrences of these clues and to make use of them in assigning labels to the different information types associated with each token.

In order to describe different kinds of separators and their functions, five operands are defined. Table 2 shows these five operands and gives examples of how they may be

used. Here <cat> refers to the information types in Table 1, and <sym>[3] is a symbol that can be used as a separator for this specific information type.

Operand	Definition	Example
<cat> *InPlaceOf* <sym>	Used as a shortcut for an information type	headword *InPlaceOf* ~
<cat> *StartsWith* <sym>	Information type begins with this separator	pronunciation *StartsWith* [
<cat> *EndsWith* <sym>	Information type ends with this separator	translation *EndsWith* ;
<cat> *PreviousEndsWith* <sym>	Previous information type ends with this separator	translation *PreviousEndsWith*
<cat> *Contains* <sym>	Information type contains this separator	derived *Contains* .

Table 2: Operands used to model separators

The tagging algorithm proceeds as follows. First the entry is divided into segments using the font styles and separators. A *segment* is a token or a group of tokens that has the same font style or consists of given keywords and/or is separated by separators from other segments, and in practice corresponds to a single word or phrase. Each segment is assigned a single information type at the end. Since the uncertainty in the document image analysis process leads to errors in the segmentation, several rules can be created for each information type, thus allowing for a relaxed matching of OCRed output to information types. For instance, there are some cases where the separators are recognized incorrectly so we may say the pronunciation begins with either '(' or '['.

Once the entry is divided into segments, the tagging process associates a single tag with each segment. This process makes use of font styles, keywords, and separators.

As an illustration, a small subset of the resulting lexicon for the French-English dictionary given in Figure 1 is shown in Figure 3.

3.2 Stochastic Method

Unlike the rule-based method, our alternative stochastic method does not require each information type to be defined precisely and explicitly by a human operator. As before,

[3]This does not necessarily need to be a character value.

bravache	(*Headword*)	
	Pronunciation	bravaJ
	POS	masculine noun
	Translation	bully
	Translation	swaggerer
	POS	adjective
	Translation	blustering
	Translation	awaggering
bravade	(*Derived word*)	
	Pronunciation	-vad
	Gender	feminine
	Translation	bravado
	Translation	bluster

Figure 3: Sample Output Lexicon

the goal is to determine the tag of each token in an entry. If an entry is treated as a sequence of tokens, it resembles the decoding task in standard Hidden Markov Model (HMM) approaches, where the observation states correspond to the tokens in a lexical entry and the hidden states correspond to the information types associated with those tokens.

We use a standard Viterbi decoding algorithm [24] which determines the highest likelihood of a given state based on the *entire* input sequence. In order to apply this algorithm, the HMM must first be trained on enough data to induce probability matrices. We used DeMenthon and Vuilleumier's [7] HMM package. This software facilitates the implementation of entry tagging for two reasons: (1) Observations are encoded as vectors, thus allowing for the representation of several features at once; (2) Training is set up to accommodate multiple observation sequences—an important property because we can use the whole dictionary as our training set.

We use a hybrid method that combines the Baum-Welch algorithm [2] with a segmental k-means algorithm [11, 20]. This method finds local maxima by applying ten iterations of the slower Baum-Welch algorithm; then the final (smaller) hill-climbing steps of the faster segmental k-means algorithm are applied until there is no improvement, or until the system converges.

The observation sequence (or observation *vector*) used in our HMM-based approach

consists of a set of 7 features corresponding to each token of the dictionary entry: (1) *CONTENT*; (2) *FONT*; (3) *STARTING SYMBOL*; (4) *ENDING SYMBOL*; (5) *SECOND ENDING SYMBOL*; (6) *IS-FIRST*; and (7) *IS-LATIN*. *CONTENT* is associated with one of three values: Information type if the token is a keyword; *SYM* if the token is a symbol; *NUM* if the token consists only of numeric characters; otherwise, the value 1. *FONT* is the font style (normal, bold, italic) of the token. *STARTING SYMBOL* indicates whether the token is a special punctuation symbol: *ENDING SYMBOL* and *SECOND ENDING SYMBOL* indicate whether the last and second-to-last characters of the token are punctuation symbols, respectively. *IS-FIRST* indicates whether this is the first token of an entry (a boolean value). Finally, *IS-LATIN* corresponds to whether the characters in the token are Latin based characters or not.

Each token in the dictionary is transformed into an observation vector before the HMM is run. For example, the POS specification *adj.* is transformed into the observation vector '[POS Italic null . null null TRUE].' The observation vectors are provided as training data for the HMM; the Viterbi algorithm is then applied to find the most probable state sequence for the given input. There is a one-to-one mapping from the observation vectors of tokens to the states of this sequence.

The mapping of the states to information types is done using a small training sample from the dictionary. Around 400 randomly selected tokens are manually tagged. In order to find the information types corresponding to the states, we count the number of manually assigned information types that fall into each state and assign the information type with the highest count to the state.

3.3 Post-Processed Stochastic Method

When we analyzed the results of the stochastic method, we discovered that, although the results of tagging of information types are comparable to those of the rule-based approach, the identification of phrases is not as robust as that of the rule-based approach. In order to increase the performance of phrase identification, we post-process the results of the

stochastic method using keywords and separators in the dictionary.[4] The post-processing proceeds as follows: If two consecutive tokens in an dictionary entry are tagged with the same information type and if there is no separator at the end of the first token or at the beginning of the second token, we mark these two tokens as a phrase.

4 Experiments

We conducted three experiments. The first measures *Dictionary Adequacy*, the degree to which three printed, bilingual dictionaries are adequately captured by our system. The second examines *Low-Density Adequacy*, the degree of dictionary adequacy with respect to a low-density language (Cebuano). The last experiment examines the the coverage of our lexicon with respect to an automated word-for-word replacement scheme, i.e., *MT Comprehensiveness* experiment.

4.1 Dictionary Adequacy: French-English, English-Turkish, Hindi-English

We ran our three methods on three of the dictionaries from Figure 1: French-English (FE) [22], English-Turkish (ET) [1], and Hindi-English (HE) [14]. These dictionaries have different characteristics which affect the noise rate of OCR. In the FE dictionary, font is a very important feature, whereas in ET dictionary font is less important, but still necessary. In the HE dictionary, font is entirely unimportant.

We use standard precision, recall and F-measures[5] to measure the adequacy of our resulting FE, ET, HE dictionaries with respect to ground-truth data generated manually for 5 random pages of FE dictionary, 5 random pages of ET dictionary, and 5 pages worth of randomly selected entries of HE dictionary.

Some statistical information about these dictionaries is given in Table 3. The number of components represents the number of different values each feature can take in the observation vector, where the vector represents [<Content>, <Font>, <Starting symbol>,

[4] In post-processing, separators are defined using the *StartsWith*, *EndsWith* and *PreviousEndsWith* features.

[5] Precision (P) measures how accurately we tagged the entries while recall (R) is a measure of coverage. In F-measure (F-m) calculations, recall and precision are given equal weights.

	French-English	**English-Turkish**	**Hindi-English**
# of pages	528	1152	1083
# of entries	13537	36747	33020
# of tokens	304601	619715	744722
# of components	[11 4 6 9 7 2 1]	[10 4 6 7 5 2 1]	[12 2 6 10 7 2 2]

Table 3: Dictionary statistics

<Ending symbol>, <Second ending symbol>, <Is-first token>, <Is-Latin>]. For 5 pages of ground-truth from the FE dictionary, there are 167 entries and 2918 tokens, for the ET dictionary, there are 193 entries and 2555 tokens, and for the HE dictionary, there are 136 entries and 2808 tokens.[6]

We evaluated our entry-tagging approach on a number of complete dictionaries by comparing the results against our manually prepared ground truth. We performed two different sub-experiments. The first evaluation was word-based, where each token is viewed as a *single-word* entry, even if it is part of a phrase. The second was phrase-based, i.e., we considered *multi-token* entries to be grouped together as a logical phrase.[7]

As an example of the phrase-based evaluation, consider the FE dictionary from Figure 1. Here, the correct translation for *brasure* is *brazed seam*. If the system produces the translation 'brazed seam' (as a unit), then this is counted as a correct entry. If, on the other hand, the system produces two independent words 'brazed' and 'seam', this result is counted as incorrect. Phrase-based evaluation is important for machine translation, but word-based evaluation is also significant since certain cross-language applications (e.g., CLIR) treat all translations of a word as a list.

The results of our experiments are presented in Table 4. We tabulated percentages for two different configurations: "*all information types* (AIT)" and "*headword and derived word translations only* (HDT)". The first gives the result for all information types present

[6]It is worth noting that the derived word and usage example have translations for these dictionaries, but these translations have the same properties as the headword translation. Thus, we did not explicitly prepare rules for these two types of translations; instead, we assigned the same information type to all translations in the training data. The type of the translation is identified by the information type of the last token bearing that translation (i.e. headword, derived word, or usage example).

[7]In the phrase-based evaluation, if a multi-token entry is assigned one information type in the ground truth, we considered the tagging correct only if the same multi-token entry was assigned the same information type by the system.

French-English Dictionary

		All Information Types			Hw/Der Word Trans		
System Type	**Eval. method**	**P**	**R**	**F-m**	**P**	**R**	**F-m**
Rule-based	Word-based	72.55	72.55	72.55	67.93	**77.27**	**72.30**
Rule-based	Phrase-based	74.73	75.19	74.96	64.97	74.51	69.41
Stochastic	Word-based	**77.62**	**77.62**	**77.62**	70.71	62.47	66.34
Stochastic	Phrase-based	55.78	69.97	62.08	48.15	54.72	51.23
Post-pr. st.	Word-based	**77.62**	**77.62**	**77.62**	**76.65**	67.72	71.91
Post-pr. st.	Phrase-based	67.59	72.86	70.13	74.46	67.32	70.71

English-Turkish Dictionary

		All Information Types			Hw/Der Word Trans		
System Type	**Eval. method**	**P**	**R**	**F-m**	**P**	**R**	**F-m**
Rule-based	Word-based	86.97	86.97	86.97	**84.77**	87.93	86.33
Rule-based	Phrase-based	**89.04**	87.93	**88.48**	84.01	89.22	86.53
Stochastic	Word-based	88.14	**88.14**	88.14	80.09	85.91	82.90
Stochastic	Phrase-based	40.03	62.86	48.91	17.24	39.14	23.94
Post-pr. St.	Word-based	88.14	**88.14**	88.14	84.22	**90.33**	**87.17**
Post-pr. St.	Phrase-based	84.55	85.10	84.83	82.25	87.59	84.84

Hindi-English Dictionary

		All Information Types			Hw/Der Word Trans		
System Type	**Eval. method**	**P**	**R**	**F-m**	**P**	**R**	**F-m**
Rule-based	Word-based	85.93	**85.93**	**85.93**	**78.64**	**78.25**	**78.44**
Rule-based	Phrase-based	**85.99**	85.07	85.53	74.16	78.03	76.04
Stochastic	Word-based	72.69	72.69	72.69	45.87	53.15	49.24
Stochastic	Phrase-based	51.62	50.45	51.03	23.79	17.85	20.39
Post-pr. St.	Word-Based	72.69	72.69	72.69	46.93	54.37	50.38
Post-pr. St.	Phrase-based	56.69	64.55	60.37	37.91	50.86	43.44

Table 4: Experiment Results

in the dictionary. The second considers only headword and derived word translations. The results specify an average value over the ground truth for each dictionary.

When the font is a distinguishing feature, as in FE and ET, the stochastic method usually outperforms the rule-based method. However, the rule-based method outperforms stochastic method if the font is not a distinguishing feature, such as in the HE dictionary. Moreover, the stochastic method alone is not very successful in identifying phrases regardless of the structure of the dictionary. The post-processing stochastic method improves the F-measure of the phrase-based results between 13-73% when AIT are considered, and between 38-254% when HDT are considered. Therefore, for dictionaries that contain phrases, post-processing is necessary when the stochastic method is used.

4.2 Low-Density Adequacy: Cebuano-English

We evaluated a Cebuano-English [5] dictionary using a different approach. For this dictionary, we investigated the handling of the POS, Cebuano and English terms. We use 100 randomly selected (ground-truth) entries from the original dictionary as the basis of our comparison against the generated lexicon. Our evaluation involves a verification of only these information types; each token was categorized as one of three types: (1) missing—not in the generated lexicon; (2) extra—not in the original dictionary; (3) incorrect—tagged correctly, but incorrect because of OCR noise. Table 5 presents our results. In addition, we found out that among the correct Cebuano terms, 12.89% of them has incorrect accents because of OCR noise.

	Cebuano	**POS**	**English**
Correct	95.36	95.00	88.12
Missing	2.06	5.00	4.95
Extra	0.00	0.00	3.96
OCR error	2.58	0.00	2.97

Table 5: Cebuano Experiment Results

4.3 MT Comprehensiveness

To approximate the degree to which our lexicons are comprehensive enough for machine translation, we conducted an experiment involving the use of French-English lexicons produced by the rule-based technique and stochastic technique described above. We performed an automatic word-for-word English replacement of the words in the French Bible using these two lexicons, and calculated the coverage against its parallel English Bible, using the standard IR-based recall metric. Table 6 presents the recall values for the lexicons produced by the three methods. Overall recall is the recall of the whole Bible, whereas sentence recall is the average recall across independent verses. The recall results for the stochastic method are much higher, supporting our claim that for the dictionaries in which font is an important distinguishing feature (e.g., the French-English dictionary), the stochastic method generally outperforms the rule-based method.

Lexicon	Overall Recall	Sentence Recall
Rule-based lexicon	49.57	47.65
Stochastic lexicon	69.75	67.83

Table 6: MT Comprehensiveness Experiment Results

5 Conclusion and Future Work

In this paper, we proposed three methods for the solution to the problem of tagging dictionary entries in bilingual dictionaries in order to acquire an MT lexicon from printed dictionaries. The first method relies on rules and information about the structure of the dictionary from an operator. The second one is HMM-based, requiring only a very small amount of training data to determine the information types of tokens. The third one involves post-processing on the second method to improve the results for phrasal entries. We tested our system using different kinds of dictionaries including ones with non-Latin scripts, and we demonstrated that these methods give promising results, especially for low-density languages. When electronic resources are limited and the need for online dictionaries is crucial for several NLP applications, our approach is promising in that it provides rapid lexicon acquisition with minimal human assistance.

A future area to investigate is the use of more than one dictionary for the same language—as an approach to increasing recall. Finally, we plan to investigate the use of English-heavy resources to improve our results—e.g., to generate POS information (critical to the task of MT) when it is not available. This can be done by applying categorial matching of multiple English translations (for each bilingual entry) against a large POS database [10].

References

[1] Robert Avery, Serap Bezmez, Anna G. Edmonds, and Mehlika Yaylalı. *Redhouse İngilizce-Türkçe Sözlük*. Redhouse Yayınevi, 1974.

[2] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. In *Inequalities III: Proceedings of the*

Third Symposium on Inequalities, pages 1–8, University of California, Los Angeles, 1972. Academic Press.

[3] Hans C. Boas. Bilingual framenet dictionaries for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain, 2002.

[4] Nicoletta Calzolari and Allessandro Lenci. Computational lexicons for cross-language information retrieval. In *Cross-Language Information Retrieval: A Research Roadmap Workshop at SIGIR-2002*, Tampere Finland, 2002.

[5] Jan Edmund Carlsen. *Cebuano-English-Swedish Lexikon*. JEC Computext, 1999.

[6] Ann Copestake, Ted Briscoe, Piek Vossen, Alicia Ageno, Irene Castellon, Francesc Ribas, German Rigau, Horacio Rodríguez, and Anna Samiotou. Acquisition of lexical translation relations from mrds. *Machine Translation*, 9(3–4):183–219, 1995.

[7] Daniel DeMenthon and Marc Vuilleumier. Lamp_hmm v.0.9. http://www.cfar.umd.edu/-~daniel/LAMP_HMM.zip, 2003. software.

[8] Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 414–420, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[9] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, California, June 1991.

[10] Nizar Habash and Bonnie J. Dorr. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, California, 2002.

[11] B.H. Juang and L.R. Rabiner. The segmental k-means algorithn for estimating parameters of hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(9):1639–1641, 1990.

[12] Huanfeng Ma and David Doermann. Bootstrapping Structured Page Segmentation. In *Proceedings of SPIE Conference Document Recognition and Retrieval*, Santa Clara, CA, January 2003.

[13] Song Mao and Tapas Kanungo. Stochastic language models for automatic acquisition of lexicons from printed bilingual dictionaries. In *Document Layout Interpretation and Its Applications*, Seattle, WA, 2001.

[14] R.S. McGregor, editor. *The Oxford Hindi-English Dictionary.* Oxford University Press, 1993.

[15] I. Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, June 2000.

[16] Arul Menezes and Stephen D. Richardson. A Best-First Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora. In *Proceedings of ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, Toulouse, France, 2001.

[17] Mary Neff and Michael McCord. Acquiring Lexical Data from Machine-Readable Dictionary Resources for Machine Translation. In *Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI-90)*, pages 85–90, Austin, Texas, 1990.

[18] M.S. Neff, B. Blaser, J.M. Lange, Hubert Lehmann, and Isabel Zapata Dominguez. Get it where you can: Acquiring and maintaining bilingual lexicons for machine translation. In *Working Notes, Building Lexicons for Machine Translation, AAAI-93 spring symposium, Technical Report SS-93-02*, page 104, Stanford, CA, 1993.

[19] Casey Palowitch and Darin Stewart. Automating the structural markup process in the conversion of print documents to electronic text. In *Digital Libraries '95: The Second Annual Conference on the Theory and Practice of Digital Libraries*, Austin, Texas, 1995.

[20] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals Of Speech Recognition.* Prentice Hall, 1993.

[21] Philip Resnik. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland, College Park, Maryland, June 1999.

[22] K. Urwin. *Langenscheidt's Standard French Dictionary.* Langenscheidt, Germany, 1988.

[23] T. Utzuro, T Horiuchi, Y. Chiba, and T. Hamamoto. Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on www news sites. In *Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, California, 2002.

[24] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. In *IEEE Transactions on Information Theory*, pages 260–269, 1967.

[25] Piek Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Dordrecht, 1998.

[26] Donald Walker and Robert Amsler. The use of machine-readable dictionaries in sublanguage analysis. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains*, pages 69–83. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.